Multitask Learning without Label Correspondences

Novi Quadrianto[1] | Alex Smola[2] | Tiberio Caetano[1] | S.V.N. Vishwanathan[3] | James Petterson[1]

1: SML-NICTA & RSISE-ANU | 2: Yahoo! Research | 3 Purdue University

Abstract

- We propose an algorithm to perform multitask learning where each task has potentially distinct label sets and label correspondences are not readily available;
- Our method directly maximizes the mutual information among the labels;
- We show that the resulting objective function can be efficiently optimized using existing algorithms;
- Our proposed approach has a direct application for data integration with different label spaces, for the purpose of classification, such as integrating Yahoo! and DMOZ web directories.

Motivating Example

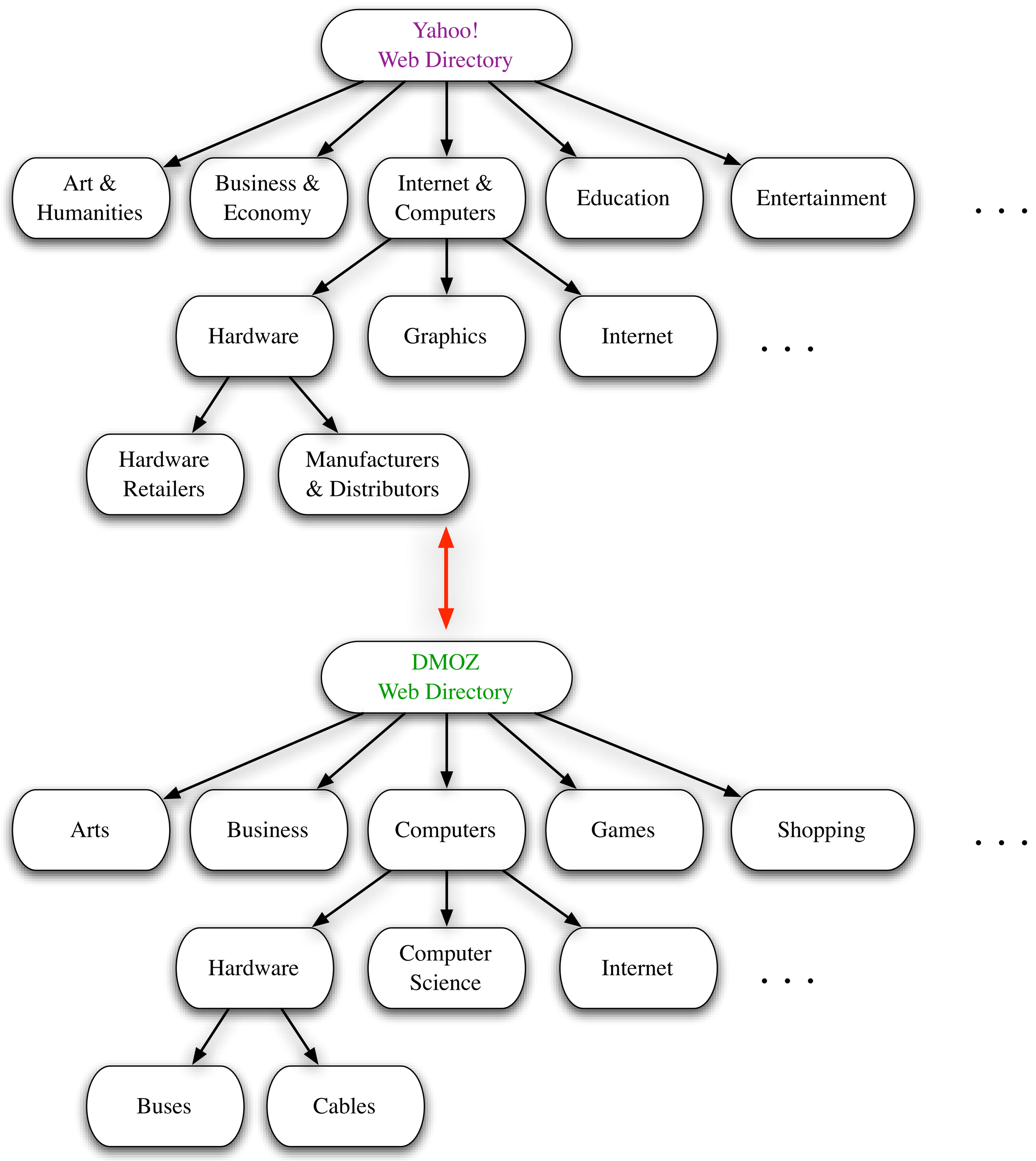
Web Ontologies Integration

Goal:

- Building a categorizer for the Yahoo! directory while taking into account other related web directories.

(Potential) Problems:

- Some section heading and sub-headings may be named differently in the two directories;
- Different editors may have made different decisions about the ontology depth and structure, leading to incompatibilities;
- Ontologies evolve with time and certain topic labels may die naturally due to lack of interest or expertise while other new topic labels may be added to the directory;
- Given the large label space, it is unrealistic to expect that a label mapping function is readily available.



Maximum Entropy Duality for Conditional Distributions

Recall the definition of the Shannon entropy, $H(y|x) := -\sum_y p(y|x) \log p(y|x)$, where $p(y|x)$ is a conditional distribution on the space of labels $\mathcal{Y}$. Let $x \in \mathcal{X}$ and assume the existence of $\phi(x,y) : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{H}$, a feature map into a Hilbert space $\mathcal{H}$. Given a data set $(X,Y) := \{(x_1,y_1),\ldots,(x_m,y_m)\}$, where $X := \{x_1,\ldots,x_m\}$, define $\mathbf{E}_{y\sim p(y|X)}[\phi(X,y)] := \frac{1}{m}\sum_{i=1}^m \mathbf{E}_{y\sim p(y|x_i)}[\phi(x_i,y)]$, and $\mu = \frac{1}{m}\sum_{i=1}^m \phi(x_i,y_i)$. With the notations, we have (Altun & Smola 2006):

$$\min_{p(y|x)} \sum_{i=1}^m -H(y|x_i) \text{ s.t. } \left\|\mathbf{E}_{y\sim p(y|X)}[\phi(X,y)] - \mu\right\|_{\mathcal{H}} \le \epsilon \text{ and } \sum_{y\in\mathcal{Y}} p(y|x_i) = 1 \tag{1a}$$

$$= \max_{\theta} \langle\theta,\mu\rangle_{\mathcal{H}} - \sum_{i=1}^m \log\sum_y \exp(\langle\theta,\phi(x_i,y)\rangle) - \epsilon\|\theta\|_{\mathcal{H}}. \tag{1b}$$

Note that: by enforcing the moment matching constraint exactly, that is, setting $\epsilon = 0$, we recover the well-known duality between maximum (Shannon) entropy and maximum likelihood (ML) estimation.

Multitask Mutual Information

Problem Setting

Assume that we are given two data sources with labels $Y = \{y_1,\ldots,y_c\} \subseteq \mathcal{Y}$ and observations $X = \{x_1,\ldots,x_m\} \subseteq \mathcal{X}$ (resp. $Y' = \{y'_1,\ldots,y'_{c'}\} \subseteq \mathcal{Y}'$ and $X' = \{x'_1,\ldots,x'_{m'}\} \subseteq \mathcal{X}'$). The observations are disjoint but we assume that they are drawn from the same domain, i.e., $\mathcal{X} = \mathcal{X}'$.

Objective Function

Assumption: The labels are different yet correlated we should assume that the joint distribution $p(y,y')$ displays high mutual information ($I(y,y') = H(y) + H(y') - H(y,y')$) between y and y'.

Since the marginal distributions over the labels, $p(y)$ and $p(y')$ are fixed, maximizing mutual information can then be viewed as minimizing the joint entropy $H(y,y') = -\sum_{y,y'} p(y,y') \log p(y,y')$. This reasoning leads us to adding the joint entropy as an additional term for the objective function of the multitask problem, as follows:

$$\underset{p(y|x)}{\text{maximize}} \sum_{i=1}^m H(y|x_i) + \sum_{i=1}^{m'} H(y'|x'_i) - \lambda H(y,y') \text{ for some } \lambda > 0 \tag{2a}$$

$$\text{s.t. } \left\|\mathbf{E}_{y\sim p(y|X)}[\phi(X,y)] - \mu\right\| \le \epsilon \text{ and } \sum_{y\in\mathcal{Y}} p(y|x_i) = 1 \tag{2b}$$

$$\left\|\mathbf{E}_{y'\sim p(y'|X')}\left[\phi'(X',y')\right] - \mu'\right\| \le \epsilon' \text{ and } \sum_{y'\in\mathcal{Y}'} p(y'|x'_i) = 1. \tag{2c}$$

Difficulties (and our workaround):

- The joint entropy term $H(y,y')$ is concave, hence the above objective of the optimization problem is not concave in general (it is the difference of two concave functions). We therefore propose to solve this non-concave problem using the concave convex procedure (CCCP);
- The joint distribution between labels $p(y,y')$ is unknown. We will estimate this quantity (therefore the joint entropy quantity) from the observations x and x'. Further, we assume that y and y' are conditionally independent given an arbitrary input $x \in \mathcal{X}$, that is $p(y,y'|x) = p(y|x)p(y'|x)$. This gives the estimated quantity of $H(y,y'|X)$ (and similarly for $H(y,y'|X')$) as

$$H(y,y'|X) = -\sum_{y,y'}\left[\frac{1}{m}\sum_{i=1}^m p(y|x_i,\theta)p(y'|x_i,\theta')\right]\log\left[\frac{1}{m}\sum_{j=1}^m p(y|x_j,\theta)p(y'|x_j,\theta')\right]. \tag{3}$$

Optimization

CCCP Procedure

CCCP finds the successive linear lower bounds on $H(y,y')$ and to solve the resulting decoupled convex problems in $p(y|x)$ and $p(y'|x')$ separately. Define $g_y(x_i) := -\partial_{p(y|x_i)} H(y,y'|X)$ and similarly $g_y(x'_i)$, $g_{y'}(x_i)$, and $g_{y'}(x'_i)$ for the derivative with respect to $p(y|x'_i)$, $p(y'|x_i)$ and $p(y'|x'_i)$, respectively. This leads to the following decoupled optimization problems in $p(y|x_i)$ and an analogous problem in $p(y'|x'_i)$:

$$\min_{p(y|x)} \sum_{i=1}^m\left[-H(y|x_i) + \lambda\sum_y g_y(x_i)p(y|x_i)\right] + \sum_{i=1}^{m'}\left[-H(y|x'_i) + \lambda'\sum_y g_y(x'_i)p(y|x'_i)\right] \tag{4a}$$

$$\text{subject to } \left\|\mathbf{E}_{y\sim p(y|X)}[\phi(X,y)] - \mu\right\| \le \epsilon. \tag{4b}$$

Algorithm

Input: Datasets (X,Y) and (X',Y') with $\mathcal{Y} \neq \mathcal{Y}'$, number of iterations N
Output: θ, θ'
Initialize $p(y) = 1/|\mathcal{Y}|$ and $p(y') = 1/|\mathcal{Y}'|$
for $t = 1$ to N **do**
  Solve the dual problem of (4) w.r.t. $p(y|x,\theta)$ and obtain θ_t
  Solve the dual problem of (4) w.r.t. $p(y'|x',\theta')$ and obtain θ'_t
end for
return $\theta \leftarrow \theta_N, \theta' \leftarrow \theta'_N$

Experiments

Dataset Statistics:

- 19186 webpages for Yahoo! and 35270 for DMOZ;
- We use the standard bag-of-words representation with TF-IDF weighting as our features (#dim = 27075).

Yahoo!:

Topic	MTL/STL (% Imp.)	Topic	MTL/STL (% Imp.)
Arts	56.27/55.11 (2.10)	News & Media	15.23/14.83 (1.03)
Business & Economy	66.52/66.88 (-0.53)	Recreation	68.81/67.00 (2.70)
Computer & Internet	52.57/48.12 (9.25)	Reference	26.65/24.81 (7.42)
Education	62.48/63.02 (-0.85)	Regional	62.85/61.86 (1.60)
Entertainment	63.30/61.37 (3.14)	Science	78.58/79.75 (-1.46)
Government	24.44/22.88 (6.82)	Social Science	31.55/30.68 (2.84)
Health	85.42/85.27 (1.76)	Society & Culture	49.51/49.05 (0.94)

DMOZ:

Topic	MTL/STL (% Imp.)	Topic	MTL/STL (% Imp.)
Arts	57.52/57.84 (-0.5)	Reference	67.42/67.42 (0)
Business	54.02/53.05 (1.83)	Regional	28.59/28.56 (0.10)
Computers	75.08/75.72 (-0.8)	Science	42.67/42.09 (1.38)
Games	78.58/78.58 (0)	Shopping	75.20/74.62 (0.54)
Health	82.34/82.55 (-0.14)	Society	57.68/58.20 (-0.89)
Home	67.47/67.47 (0)	Sports	83.49/83.53 (-0.05)
News	61.70/62.01 (-0.49)	World	87.80/87.57 (0.26)
Recreation	58.04/58.25 (-0.36)		

References

- Y. Altun and A.J. Smola. Unifying divergence minimization and statistical inference via convex duality. *Proc. Annual Conf. Computational Learning Theory*, 2006.